Automatikus verselemzés tanuló algoritmusok alkalmazásával

Lesi Zoltán[1]

[1] Nokia Hungary Kft. 1092 Budapest Köztelek utca 6.
zoli@nix.hu

Kivonat: Cikkünkben a számítógépes verstani elemzés tárgykörét, valamint ezen új területen elért eredményeinket mutatjuk be. Egy automatikus vers elemző jelentősen megkönnyítheti a nyelvészek, irodalmárok munkáját, különösen nagy korpusz vizsgálata esetén. Segítheti az irodalmi és nyelvi állítások bizonyítását. A fonetikai vizsgálatok képezik a verstani analizálás alapját. Részletesebben a metrika, alliteráció és rím gépi meghatározásával foglalkozunk, melyhez tanuló algoritmusokat alkalmazunk. Eredményeinket TEI P4 konzorciumnak megfelelő XML-lel írjuk le, mely tartalmazza a versek fonetikai, szótagszerkezeti és metrikai leírását, végrímeket, alliterációkat és a szófaji elemzést is. Az algoritmusunkat Weöres Sándor szonettjein és szonett fordításain teszteltük, igazolva elképzeléseink helyességét.

1. Bevezetés

A szépirodalmi szövegeket többféle szempont szerint csoportosíthatjuk. Horváth Iván *A vers* [3] című könyvében három megközelítést mutat be. Ha a transzcendens versolvasó szemüveget tesszük fel, a verset nem tudjuk másnak látni, mint költeménynek, ilyenkor a verselmélet egybeolvad a költészet elméletével. Tekinthetjük a verset nyelvi egyetemességnek: olyan megjelölt beszédmódnak, amely valamiképpen minden természetes nyelvben létezik. A harmadik módszer szerint a vers az, amit egy bizonyos irodalmi hagyomány részesei annak tartanak.

A számítógépes nyelvészetben a verselemzés új kutatási terület, hiszen magyar nyelvű szövegekre kidolgozott (vagy magyar szerző munkájaként ismert) automatikus verselemző programról nincs tudomásunk. A probléma megoldása talán azért is váratott magára, mert nem könnyű elhatárolni a megoldható és egyelőre megoldhatatlannak tűnő részeket.

Fónagy Iván nemzetközi hírű nyelvész, pszichológus a hatvanas években megtervezetett [2] egy programot, amely prózai és verses szövegekkel foglalkozik, majd kiegészítette két másik fejezettel: „Program köznyelvi szövegekre", „Program költői szövegekre". A tervek világos, pontokba szedett szempontokat tartalmaznak, amelyek statisztikai jellegű információkra mutatnak. Ez a szempontrendszer adta a nyelvészeti és verstani alapot programunkhoz.

1.1 Weöres Sándor és a korpusz

Weöres Sándor (1913-1989) költő, műfordító, író. Bori Imre tanulmányában [1] írja, hogy jellemző Weöres költészetére egy fontos zenei mozzanat a kettősség: ha Weöres verseinek nagyobbik hányadát a zene fogalmával helyettesíthetjük, úgy van egy verscsoportja, amely a „nem-zenét" jelenti. Ez a zenei elv összefogja Weöres költőiségének alapvető törekvését, amely a harmónia utáni vágyban s a költői megvalósulásában nyilatkozik meg. Weöres szerint: „A szonett első nyolc sora a nyolcoldalú kristály, az oktaéder: a végső hat sor az előbbiek ismétlése, más összeállításban, más összefüggésekkel."[10]

Nagy L. János és Alexin Zoltán 1999 és 2002 között létrehozták a virtuális kritikai kiadás 'editio princeps'-ét, hogy minél teljesebb Weöres-korpuszon dolgozhassanak. [6] Az 1986-ban megjelent háromkötetes „Egybegyűjtött írások" című gyűjteményt vették alapul.

2. Automatikus fonetikai elemzés

A szövegek fonetikus konverziójához Kassai Ilona fonetikus átíró szabályokat tartalmazó táblázatát alkalmaztuk. [4] Megvizsgálva a karakter környezetét, szintaktikai elemzés alapján döntjük el, hogy az adott helyen digramma van-e. Összetett szavak határán előfordulhat olyan eset, mikor a látszólag összetartozó karaktereket különböző hangokká kell alakítani (pl. százszor).

Amennyiben egy magyar szóról van szó, akkor a fonologikus szabályokat, idegen szó esetén az „idegen szavak átírási szótárát" vesszük figyelembe. A fonetikai átírás másik problémája, hogy az idegen nevek, szavak más átírást követelnek, mint a magyarok. Tovább nehezíti a helyzetet, hogy ezek többnyire magányosan szerepelnek. A problémát nyelvazonosító rendszer [5] alkalmazásával sikerült megoldani. Különféleképpen jelöljük az allofónokat, hangváltozatokat.

3. Automatikus verstani elemző

A morfológiai elemzést (melyet a versek HuMor [8] elemzéséből nyerünk) és a hangtani vizsgálatokat felhasználva meghatározhatóak a számunka fontos alkalmazások a metrika, az alliterációk és a rímek. A megalapozott végeredményeket egy XML fájlban összegezzük, és lekérdezéseket hajtunk végre.

A sorvégi egybecsengések illetve alliterációk vizsgálatához szükség van a hangok összehasonlítására. A fónagyi tervezet tartalmaz egy leírást a fonémák eltérési fokáról, mely egyszerűen algoritmizálható.

A metrika, az alliteráció és a sorvégi egybecsengések alkalmazásait heurisztikus és tanuló algoritmusokkal is meghatároztuk, majd a részeredményeket szavazással egyesítettük.

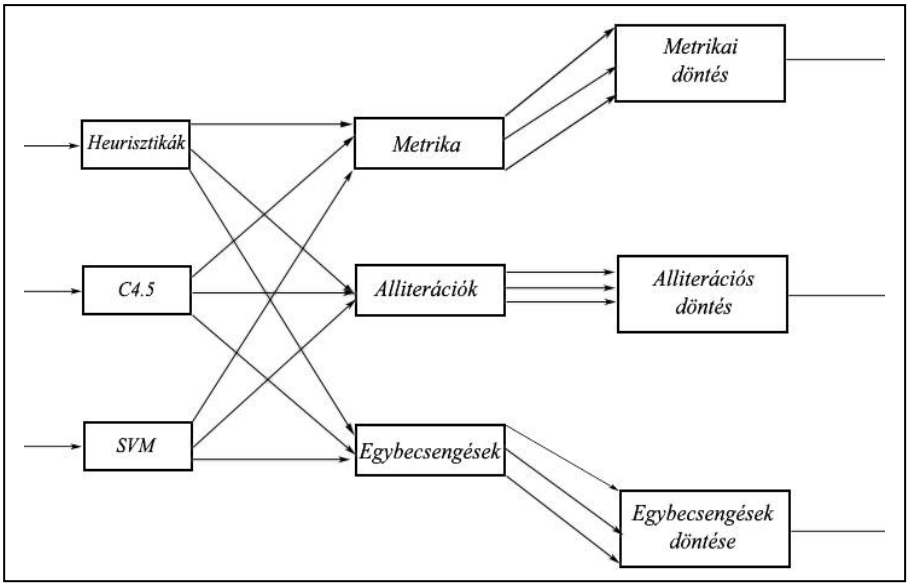


Fig. 2. A heurisztikák és a tanuló algoritmusok együttműködése

A következő táblázat a 152 szonettből álló korpuszra lefuttatott verselemző program eredményeit mutatja be:

	Heurisztikák	C4.5	SVM	Döntés után
Metrika	100%	100%	99.87%	100%
Alliterációk	97.36%	97.30%	90.92%	99.59%
Sorvégi egybecsengések	53.75%	81.54%	78.94%	81.73%

4. Az elkészült verselemző rendszer

A 2004. októberében elkezdett tudományos munka legfőbb eredménye, hogy a Fónagy-tervezet fejezetei alapján megterveztük és implementáltuk az első magyar automatikus verselemzőt. Az első időszak feladatai: a Fónagy-tervezet értelmezése; megvalósíthatósági tanulmány, DTD tervezet készítése, valamint a verstani, hangtani ismeretanyag összegyűjtése. A fónagyi tervezetben a statisztikai szempontokat visszavezettük alapadatokra. Nagy L. János kurzusán teszteltük a szempontrendszert, amely később a program alapjává vált. A program ellenpontjaként a diákok elemezték a verseket.

Összegyűjtöttük a virtuális kritikai kiadásból Weöres Sándor összes (101) szonettjét, valamint 91 szonett fordítását. Annak ellenére ragaszkodtunk a szonett formához, hogy a korpuszunk így aránylag kis méretű lett. Mindenképpen homogén versformájú anyagot akartunk: szonettekből találtunk a legtöbbet. Az egyes versekre

adott eredmény könnyebben összevethető a korpuszéval, és a szonett formát is jellemzi.

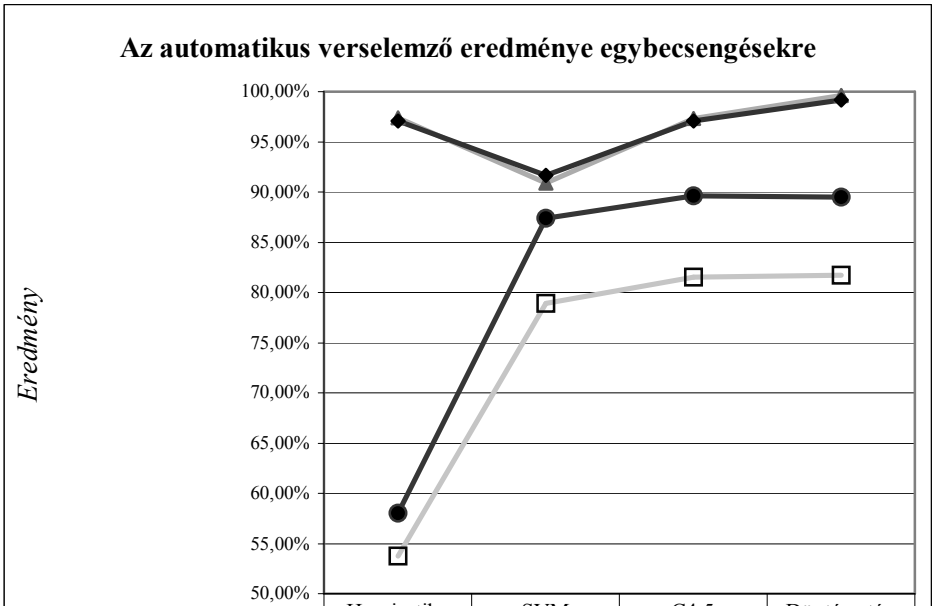
A fonetikai és a fonológiai elemző alapja a fonetikai átíró szabályrendszer. A további vizsgálataink miatt pontos fonetikai eredmények szükségeltettek. A digrammák szétvágását (pl. százszor) morfológiai elemzés alapján végeztük, tesztjeink igazolják a módszer helyességét. A versekben előforduló idegen szavak elszigetelten fordulnak elő, felismerésükhöz nyelvazonosító rendszert használtunk. Bizonyos hangoknak több ejtésváltozata (allofónja) van, így néhány újabb szabállyal kellett bővíteni a rendszert, valamint a „méh" típusú *h* miatt a szóvégmutató szótárt [7] alkalmaztuk. A hangok egymásra hatásakor megváltozik a fonetikus leírás, ezért fonologikus szabályokat alkalmazunk.

A magánhangzók eltérési fokának meghatározása hiányzott a Fónagy-tervezetből, ezt „A mássalhangzók eltérési foka" című fejezet alapján pótoltuk. A metrika, az alliterációk és a rímek heurisztikus meghatározása a Fónagy-tervezet alapján történt.

Az eredmények pontosításához C4.5 és SVM tanuló algoritmusokat használtuk. Létrehoztuk a három alkalmazás(metrika, alliteráció és sorvégi egybecsengések) modelljét, és dekomponáltuk a feladatokat. A metrikánál a szótag hosszúságát, az alliterációnál a kezdőbetűk egybecsengését elemeztük. A morfológiai elemzés segítségével a kötőszavakat és a névelőket kiszűrtük, mert csak a szűkebb értelemben vett alliterációkat kerestük. A rímeket szétbontottuk egy szótagos egybecsengésekre, ahol elsősorban a szótag tulajdonságait, a tanulóhalmazban pedig az egymással rímelő sorpárok gyakoriságát vizsgáltuk.

A két tanuló algoritmus és a heurisztika eredményeit szavazással egyesítettük, a végeredmény a sorvégi egybecsengések vizsgálatakor pontosabb, a másik két alkalmazásban pedig mind a tanuló modellek, mind a heurisztikus algoritmusok kiváló eredményt adtak, ezért érdemes volt tanuló algoritmusokat használni. Meglepő eredmény, hogy a C4.5 pontosabban osztályoz, mint az SVM, ezt a két szintű osztályozás és a diszkrét értékek okozzák. Biztosak lehetünk abban, hogy más feladatokra, más adatokon az SVM lenne a megfelelőbb.

A következő diagram a 152 szonettből álló tesztkorpuszra, és az *Átváltozások* ciklusra lefuttatott sorvégi egybecsengéseket, alliterációkat vizsgáló alkalmazások eredményeit mutatja be.



	Heurisztika	SVM	C4.5	Döntés után
Sorvégi egybecsengések 152 szonettre	53,78%	78,94%	81,54%	81,73%
Sorvégi egybecsengések az "Átváltozások" 40 szonettjére	58,01%	87,39%	89,61%	89,47%
Alliterációk 152 szonettre	97,36%	90,92%	97,30%	99,59%
Alliterációk az "Átváltozások" 40 szonettjére	97,05%	91,64%	97,05%	99,20%

Fig. 3. Az automatikus verselemző eredménye a két vizsgált korpuszra

A végeredményeket egy nemzetközileg elismert TEI kompatibilis XML fájlban összegeztük [9]. ehhez felhasználtuk a már kész DTD terveket. A TEI ajánlást betartva, minimális módosítással, oldottuk meg a konverziót. A DTD a TEI ajánláshoz képest a következő új elemeket tartalmazza: <w>, <ana>, <cons>, <c>, <link>, <linkGrp>, valamint bővíteni kellett a <body>, <lg>, <l> elemeket is.

Az elemzés elvégzése után – a Fónagy-tervezet alapján – lekérdezéseket hajtottunk végre. Néhány fonetika, metrika, alliteráció és sorvégi egybecsengések lekérdezéseit megvalósítottunk.

Magyarországon eddig nem készült automatikus verselemző szoftver. Automatizálási programunk célja, hogy a magyar nyelvű gépi verselemzés kutatását elindítsuk, és a nemzetközi tudományos áramlatba bekerüljünk.

5. A számítógépes verselemzés távlatai

A fonetikai rendszerben, komoly gondot okozott az idegen szavak átírása. Hiba már a nyelvazonosításnál felléphet, ha ismeretlen nyelvvel találkozik, mert az átírását nem biztos, hogy ismerjük. Hibás fonetikus átírás esetén, a fonemikus jegyekre, szótagszerkezetekre, digrammákra, szóhosszúságokra, rímekre, alliterációkra és metrikára pontatlan eredményeket kapunk.

A szintaktikai elemzés rohamosan fejlődik, de pontatlansága problémát okoz a szófajok, alliterációk vizsgálatakor. A Fónagy tervezetben szereplő szempontok automatizálhatóak, azonban többértelműségek miatt pontatlanságra számíthatunk. Például: A központozás hiánya bizonytalanná teheti a tagmondatokra bontást, a felsorolások, közbeékelt mondatok felismerését, a mondatok hosszát és a modalitását.

A versekben nem jelölt metrikai kétértelműségek elemzése a gép számára megoldhatatlan. Pl. „A mint B" (W.S: Önéletrajz) Az „A" hosszan ejtendő, de ezt nem jelöli semmi a gép számára. A versrendszer (időmértékes, ütemhangsúlyos vagy szimultán) felismerése is probléma lehet a gépi elemzés számára – mivel megfelel a szabályoknak – az *Átváltozások* ciklusban szereplő *A nyüzsgés* című szonett jambikus, de valójában nem is időmértékes.

A rímképletek meghatározása a rímelő sorok ismeretében előfordulhat, hogy a két sor között létezik egybecsengés, de ez a képletben nem jelenik meg.

A formai jegyek szerepét, illetve a rímek és az alliterációk jelentőségét – a mű értelmezésének tükrében – nem tudjuk számítógéppel meghatározni.

Bibliográfia

1. Bori Imre: A szintézisteremtő. In: Bori Imre huszonöt tanulmánya a XX. Századi magyar irodalomról, Forum Kiado, Újvidék, 1984.
2. Fónagy Iván: Program prózai és verses szövegek elemzéséhez. Kézirat. 1966. Javított változat: Antony, 1997.
3. Horváth Iván: A Vers. Gondolat Kiadó, Budapest, 1990.
4. Kassai Ilona: Fonetika. Nemzeti Tankönyvkiadó, Budapest, 1998.
5. Kiss Géza, Németh Géza: Skálázható szöveg-alapú nyelvazonosító módszer beszédszintézis céljára, In: III. Magyar Számítógépes Nyelvészeti Konferencia 2005. SZTE TTK Informatikai Tanszékcsoport, Szeged. Szerk.: Alexin Zoltán, Csendes Dóra.
6. Nagy L. János, Alexin Zoltán: Weöres költői nyelvének számítógépes feldolgozása. In: II. Magyar Számítógépes Nyelvészeti Konferencia 2004. SZTE TTK Informatikai Tanszékcsoport, Szeged. Szerk.: Alexin Zoltán, Csendes Dóra.
7. Papp Ferenc: Szóvégmutató szótár. Akadémiai Kiadó, Budapest. 1966.
8. Gábor Prószéky, and Balázs Kis, 1999. A Unification-based Approach to Morpho-syntactic Parsing of Agglutinative and Other (Highly) Inflectional Languages. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, 261-268. College Park, Maryland, USA
9. The Text Encoding Initiative Consortium. (http://www.tei-c.org)
10. Weöres Sándor: Oktaéder-kristály. In: Új Írás, 1977. 4